Topics as Contextual Indicators for Word Choice in SMS Conversations

Ute Winter[1], Roni Ben-Aharon, Daniel Chernobrov, Ron M Hecht[1]

[1]GM Advanced Technical Center, HaManofim Street 11, Herzeliya 46725, Israel

ute.winter@gm.com, r.ben.aharon@gmail.com,

daniel-cher@hotmail.com, ron.hecht@gm.com

Abstract

SMS dictation by voice is becoming a viable alternative providing a convenient method for texting in a variety of environments. Contextual knowledge should be used to improve performance. We propose to add topic knowledge as part of the contextual awareness of both texting partners during SMS conversations. Topics can be used for speech applications, if the relation between the conversed topics and the choice of words in SMS dialogs is measurable. In this study, we collected an SMS corpus, developed a topic annotation scheme, and built a topic hierarchy in a tree structure. We validated our topic assignments and tree structure by the Agglomerative Information Bottleneck method, which also proved the measurability of the interrelation between topics and wording. To quantify this relation we propose a naïve classification method based on the calculation of topic distinctive word lists and compare the classifiers' topic recognition capabilities for SMS dialogs with unigram language models. The results demonstrate that the relation between topic and wording is significant and can be integrated into SMS dictation.

1 Introduction

One of the largest growth areas in communication is the Short Message Service (SMS) or text messaging, as it is more popularly known. SMS grew out of what was initially a by-product of the mobile phone industry (Agar, 2003; Goggin, 2006). In fact, by 2009 text messaging has become the most frequently used communication means among teens in the US, supported by the mobile phone industry offering unlimited texting plans (Lenhart et. al., 2010).

For many reasons, voice enabled texting has become a desirable alternative in a variety of mobile scenarios. The number of speech applications for mobile phones including texting by voice is constantly growing. However, the challenges for SMS dictation by voice are multifold, from particular noise conditions, to the use of vocabulary and domain specific language, the dialogical nature of text messaging (Thurlow and Poff, 2009), and to error correction of imperfect recognition results.

Achieving a high and robust performance is crucial for the success of the application. For this purpose additional contextual factors can be integrated into the recognition process. One possible factor, the conversed topic, has influence on the speaker's choice of words. Hence, it is an important contextual factor for the prediction of the speaker's wording, since it originates in the speaker's mental concepts during a dialog situation, which is the nature of texting.

To date, research on text messaging has primarily examined socio-linguistic phenomena (e.g., Thurlow, 2003). With respect to language and communication, text messaging is still an underexamined research area. Thurlow and Poff (2009) provide a comprehensive overview of existing literature about SMS in linguistics. Moreover, there exists noteworthy work on SMS text normalization (Aw et. al., 2006; Fairon and Paumier, 2006; Cook and Stevenson, 2009; Kobus et. al., 2008; Pennell and Liu, 2010), for instance for the purpose of Machine Translation, Text-to-Speech engines or spell checking, work on SMS based question answering

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 185–193,
Portland, Oregon, June 17-18, 2011.

services (Kothari, 2009), and work on predefined SMS replies in automobiles (Wu et. al., 2010). However, conversed topics in the context of SMS discourse have not been examined in the literature, neither in linguistics nor for any Natural Language Processing applications.

Hence, in this paper we have developed a new approach to make topics useful as context knowledge for SMS dictation by voice. We describe topic annotation of a novel SMS corpus and study the influence which SMS dialog topics may have on the choice of words. Based on the results, we are able to estimate and initially quantify its impact. This research can serve as the basis for developing algorithms that use topic knowledge for SMS dictation in speech applications.

2 Topic Annotation for SMS

2.1 SMS Corpus in US English

SMS data was collected from 250 participants who conversed with another 900. Participants were distributed almost evenly across gender, two age groups, and four US regions. Participants under 30 years comprised 48% of the dataset, and participants over 30 years comprised 52% of the dataset. Within each of these two age groups, there were equal number of men and women. The demographic spread contained datasets from participants from the various regions in the USA: east coast 19%, west coast 24%, central 29%, and south 28%.

The corpus dataset contains a total number of more than 51,000 messages, chosen randomly from a significantly larger set of data, for which participants provided authentic SMS conversations from their mobile phones to online SMS backup services. Besides demographic constraints, all text messages are part of SMS conversations, each composed at least by one message and a textual response, to preserve a contextual authentic situation. A conversation is considered to be ended if a time frame of 4 hours elapses without a response. The average length of SMS conversations in the corpus is between 8-9 messages, distributed over a notably higher number of shorter conversions than longer dialogs. Altogether the corpus contains more than 5800 conversations.

Personal information of the SMS conversations was removed. Nonetheless the corpus itself is currently not published, because identifying information can be indirectly present in SMS dialogs.

The SMS corpus is semi-automatically normalized following a general guideline to transform each texted message into one which could be dictated by the user. For all following research the normalized rather than the raw SMS textual utterances are used.

Table 1 shows representative examples for text normalization.

Raw	**Normalized**
Yea b workin for hospice	yeah be working for hospice
I am at vetran @at@8 am	I am at Veteran at eight ei-em
Lets go 2 eat	Let's go to eat
You wanna go to da b walk or sumthin?	You wanna go to the bee walk or something?

Table 1: Text messages in raw and normalized format.

2.2 Topic Annotation Method

A key point for usefulness of an annotated corpus is the abstraction which maps SMS conversations present in the corpus to an abstract model serving the research goals (Wallis and Nelson, 2001; Mc Enery et. al., 2006). In our research, the corpus shall be used to explore to what extent the knowledge of one or more discussed topics, for which both SMS dialog partners try to make progress, can contribute to the performance of a speech recognition engine, where we expect the engine to be based on Statistical Language Models (SLM). Consequently, the annotation needs to enable us to trace a path from discussed topics to the choice of words and phrases in SMS conversations. This abstraction leads to our definition of the term topic and to guidelines for the annotation which are identified to be essential, when incorporating topics into speech recognition.

Other than an agreement on "what is being talked about", the definition of topic in linguistics is a matter of viewpoint and dispute (Levinson, 1983; Li and Thompson, 1976; Chafe, 1976; Molnár, 1993; Stutterheim, 1997). Moreover, a literature review has not revealed existing topic annotations which can be used for our purpose (Mc Enery et. al., 2006; Meyer, 2002). Since the inten-

tion is to build a task driven, problem oriented annotation scheme we further specify a discourse topic as observable content or story line which discourse partners follow up in an SMS conversation. Hence, we understand a topic foremost as an attribute of an SMS dialog rather than of a single SMS, or of a phrase within the dialog. We assign at least one topic to each dialog. Since dialogs can in fact contain several distinct topics, we assign all explicitly mentioned topics to a conversation and mark separately all SMS which belong doubtlessly to each topic in the context of the conversation,

Topics describe the content only, not any other level of discourse. The example in figure 1 shows a conversation with the topic 'meeting arrangement'.

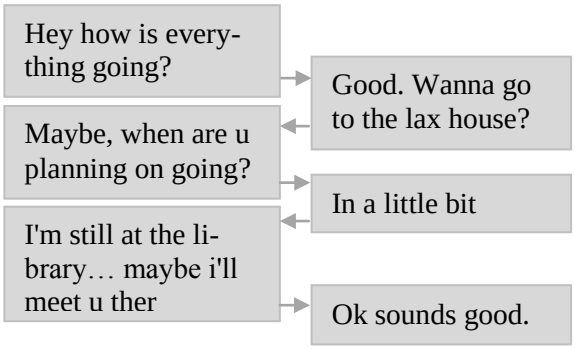

Figure 1: Example of SMS dialog about "meeting arrangement".

2.3 Topic Annotation Procedure

Discourse topics are highly domain dependent in their nature and may differ from the SMS domain to other domains, even to computer mediated communication services, like e-mail, Twitter, or Instant Messaging. Because of that, the list of SMS relevant topics evolves from the data itself. Additionally the list of possible topics always remains an open tag list, although one can expect recurring topics after a while with sparse extension of an existing topic list. Hence, the approach for annotating the SMS corpus must be manual. For this purpose a team of four annotators marked the conversations with the help of an annotation tool developed specifically for the topic annotation. To ensure annotator agreement a linguist verified and confirmed the growing topic list and all topic assignments in several iterations. Further annotation of a larger corpus may be semi-automated based on the achieved topic list.

Assigning topics to a dialog remains intuitive to a certain extent, because any mutual understanding of the dialog's content and pragmatic meaning is supported by social cues, situation awareness and world knowledge of dialog partners (Levinson, 1983; Lambert and Carberry, 1992). These knowledge dimensions need to be reconstituted during the annotation process, when assigning a new topic. One criterion is to ask if the topic is distinct from other topics with regard to describing pieces of our world knowledge dimensions, e.g. scripts and events that people repeatedly experience, or subjects, they are recurrently dealing with.

Furthermore, a task driven approach demands to determine the level of specialization and detail for topics. Even if broad topics, such as "food" or "appointment", may prove themselves to be distinct and meaningful enough for speech recognition, the annotation is done to one degree more detailed. Each topic is composed by a term and one restrictive attribute which divides a major topic into more distinctive topics. Thus "appointment" appears in the corpus divided into "cancel appointment", "attending an appointment", "meeting arrangements", and other. The advantage of the annotation procedure is twofold; it leads to a list of topics, which can be depicted in a tree structure with several levels of specialization, and, even though the annotation is targeted to a special problem, there is sufficient information to make the corpus useful for a broader range of research.

3 Corpus Analysis for Topic Usage

3.1 Properties of Topics

SMS conversations may follow up on one or more topics. Multiple topic conversations may make progress on topics even in parallel, either switching topics or addressing both within the same SMS. In general, we avoid topics which are suspected to describe the intention or strategy for the conversation rather than the content. There are a few exceptions, where the topic is implicitly or explicitly present in the dialog not only on content level but also as driving force for texting, e.g. "maintain friendship/relationship" or "small talk" (see example (2) in figure 2). The border cannot be clearly drawn in these cases.

Two topic assignments require explanation. "Small talk" is used for a group of short SMS di-

alogs, for which one cannot identify a topic. One is able to understand the dialog as a short form of friendship maintenance though, where both parties achieve mutual positive feedback about their current situation, e.g. via salutation. Therefore "small talk" is expected to be of interest regarding word usage contrary to "undefined topic". The latter is assigned to all conversations, where we do not share enough knowledge about the background and situation of the texters to understand and identify the topic of the dialog (example (3) in figure 2).

1 **Missed phone call, planned schedule**

Texter 1: Hi, sorry I missed your call. I'm actually at an appointment right now.
Texter 1: I will call you about 12:45pm. Please answer, so we can finally connect, if not I will call after 17:00.
Texter 2: O.K no problem, call me when you're free :)
Texter 1: The appointment is over, I tried calling you but you didn't answer, will talk when I'm on my way home
Texter 2: Thankyou.

2 **Small talk**

Texter 1: What's up?
Texter 2: I'm good, u?
Texter 1: I'm fine, talk to you later
Texter 2: Sure :)

3 **Topic undefined**

Texter 1: df
Texter 2: what?
Texter 1: don't forget
Texter 2: Lol :-) I won't

Figure 2: SMS dialogs with (1) multiple topics, (2) small talk, and (3) undefined topic.

All in all, the corpus contains 42.1% of dialogs with one annotated topic and 46.6% with multiple topics. The remaining 11.3% of dialogs are tagged as "undefined".

3.2 Building a Topic Tree

The identification of similar or related topics in our corpus allow for grouping them together in specific topic clusters, such as "human relations", "technology", and "transportation", and represent them in a tree structure hierarchy. The assignment to a topic cluster for each topic is determined by the relation between topics, which humans define based on their world knowledge and based on the semantic meaning of the topic.

The topic tree hierarchy consists of four levels. The nodes in the first two levels build the tree structure and represent the topic clusters. Therefore they have not been used during the annotation process. Only from level three and above the topic names are assigned to the corpus and may be leaves of the tree. A forth level is used, when third level topics are frequently used in SMS dialogs and can further be divided into meaningful sub topics.

Figure 3: Topic tree branch related to "shopping".

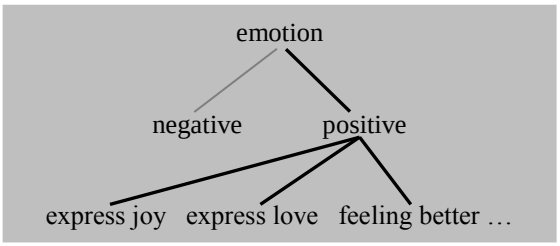

Figure 4: Topic tree branch for "positive emotion".

3.3 Topic Distribution in SMS Corpus

87.1% of all text messages are categorized in nine preferably conversed topic clusters (see figure 5), the remaining messages belong either to SMS dialogs, where the topic is labeled as undefined, or to miscellaneous, rarely conversed topics, e.g. "weather" or "religious belief".

More than 55% of all text messages are motivated by interpersonal and emotional matters. About 45% of all text messages deal with "human relations", mainly including sub topics regarding relation maintenance (36% of "human relations", e.g. "make promise", "make apology", "health condition", "small talk", a. o.), regarding relations with friends (14%), concerning relationship issues

with a partner (11%). The latter 10% converse about negative or positive emotions, nearly 50% of these dialogs expressing love. SMS dialogs from "human relations" contain 9.3 messages per dialog in the average, which is significantly more than the average of 4-6 messages in all other topic clusters.

The second most discussed topic is "activities & events" (14% of all messages), such as "going out" (32% of "activities & events" labeled messages), or "going shopping" (15%). Interestingly, the topic of "appointment & scheduling" is only the third most popular, consisting of less than 13% of all text messages.

Figure 5 shows the topic distribution in the corpus with respect to the topic tree's first hierarchy.

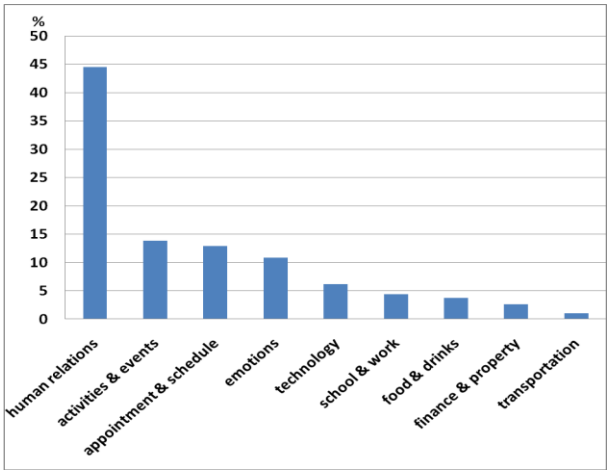

Figure 5: Topic distribution on first tree level.

Thurlow (2003) has presented a study about the communicative intent of US English text messages, describing their functional orientation rather than the content. Thurlow's findings concur in that the amount of SMS with relational and intimate orientation vs. transactional orientation is similar to the amount of SMS with interpersonal and emotional content vs. all other topic clusters.

Finally, we examine if distribution differences depend on the demographic data of the users regarding gender, age groups (18-23, 24-28, 29-35, 36-42) and regions. Users older than 42 years are not taken into account because of the limited number of text messages in the corpus.

Generally, males and females talk about the same topics in SMS conversations through all age groups and regions. However, there are still some differences between those groups worth mentioning and shown in figure 6.

While interpersonal and emotional text messages together are present in fairly equal quantity for both gender groups, females tend to express their "emotion" via text messages much more frequently than males (12.5% compared to 8.5%); likely on the expense of non-emotional "human relations" messages (46.8% for males compared to 41.9%). Furthermore, males and females have contradicting trends in "emotion" talk over ages. Females tend to express emotions more with age progression, while males have the opposite tendency. In both genders, the corpus suggests a tradeoff between the topics "human relations" and "emotion", i.e. age may change the portion of one topic on the expense of the other one.

4 Relation between Topic and Wording

4.1 Automated Validation of Topic Tree

A human annotation process is highly effective due to people's ability to exploit their mental knowledge base and mind concepts, and thus a broad range of information sources. However, even

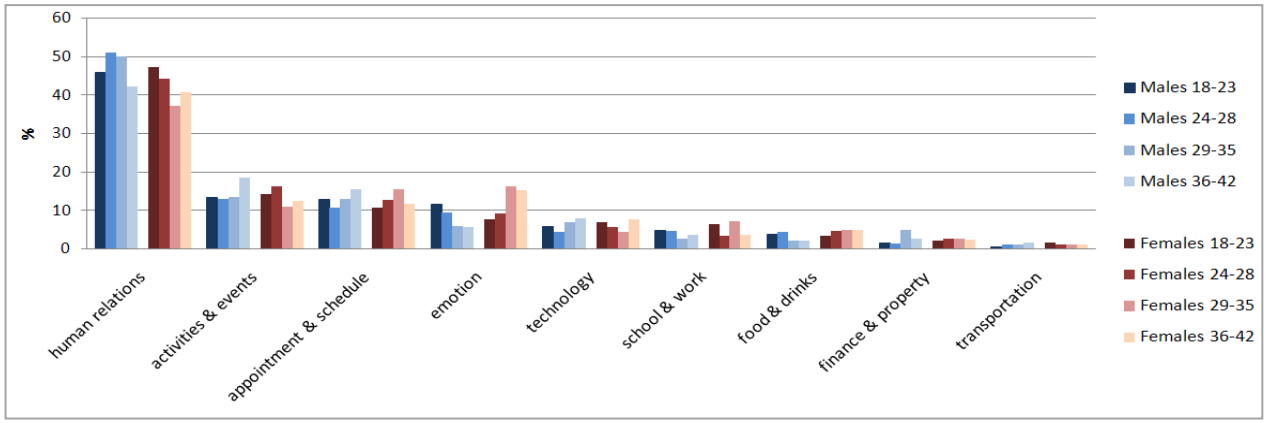

Figure 6: Topic distribution by gender (males left, females right) and age groups

in a most rigorous procedure errors may occur, especially regarding annotation and tree consistency. Therefore we need to verify the quality of the annotation. Additionally, we want to ensure that relevant algorithms can trace the interrelation between topics and the choice of words in SMS.

In order to verify both requirements, we perform an automatic validation by applying a nuance (Hecht et al., 2009) of the Agglomerative Information Bottleneck (AIB) method (Tishby et al., 1999; Slonim and Tishby, 2000). This derivative of the AIB is a hierarchical clustering algorithm, and as such, it produces a hierarchical topic tree.

The clustering starts with each lower level topic as a singleton. In an iterative process, the two closest topics are merged to form a larger topic, where the two closest topics are defined as the ones that minimize the AIB functional (Eq. 1). The process ends when all topics are merged into a single topic.

$$L\left[p\left(\hat{x}|x\right)\right] = I\left(X;\hat{X}\right) - \beta I\left(Y;\hat{X}\right) \quad (1)$$

X, Y and $\hat{X}$ are the set of topics, set of words and clustered set of topics respectively. $I(A;B)$ is the mutual information between A and B.

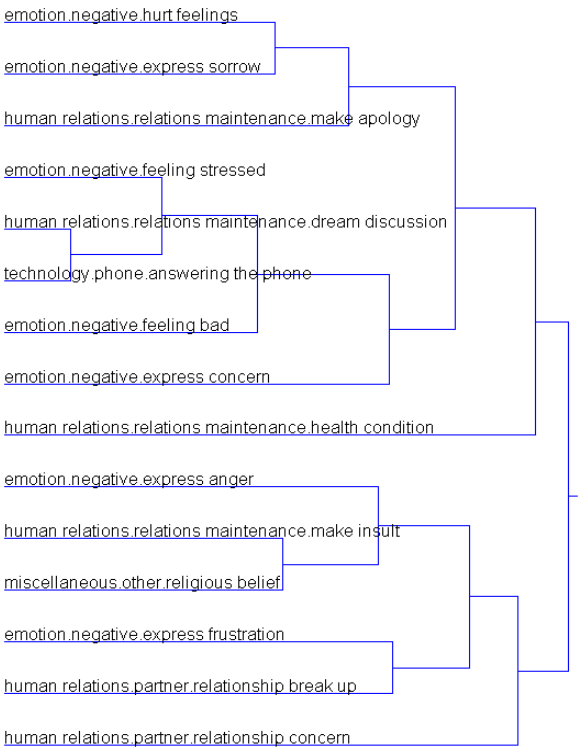

Figure 7: Tree branch of the hierarchical clustering of topics into groups.

Intuitively, the function tries to achieve two goals simultaneously. It minimizes $I\left(X;\hat{X}\right)$ which can be interpreted as finding the most compact topic representation and at the same time it maximizes $I\left(Y;\hat{X}\right)$ which can be interpreted as finding the most indicative subset of topics. These two goals contradict one another. Therefore a tradeoff parameter β is added.

Presenting the entire AIB tree is not feasible in this paper. In order to provide some intuition, a sub tree is shown in figure 7. Briefly, each AIB tree branch shows a distribution of topics that is mostly in line with the hand crafted topic tree. Even sentiments are clustered (negative sentiment for all lower level topics in figure 7), a superior achievement to the manual topic tree, where this is done only for "emotion". Moreover, it becomes evident that the interrelation between topics and wording in SMS can likely be captured automatically.

4.2 Method for Relation Discovery

Being confident regarding automatic computation, we can strive for more and aim to discover the interrelation between topics and wording in detail. Any vocabulary used in SMS dialogs can intuitively be viewed as containing information which points to one or a limited group of conversed topics, or as being general vocabulary with respect to topic distinctiveness. Such a view point entails questions. How can we extract a list of distinctive words per topic; words which are dominant in a certain topic but subordinate in others respectively? To what extent are topic distinctive words still ambiguous and are assigned to more than one topic? And ultimately, can we use topic distinctive vocabulary to recognize a list of conversed topics for each SMS dialog based on its choice of words?

Our method evolves from the questions as follows: First, we categorize the SMS vocabulary into topic distinctive vs. general vocabulary by introducing an algorithm which uses topic information as qualitative measurement to extract a list of distinctive words operating as classifiers for topics. In a second step we evaluate for each topic to what extent topic distinctive word list classifiers can recognize topics in SMS dialogs. Finally we compare the classifiers' topic recognition capabilities with unigram language models. We use only the nine first level topic clusters to guarantee that the amount of available dialogs per topic is sufficient.

4.3 Topic Distinctive Vocabulary

To categorize the vocabulary we calculate for each word w_i with at least 4 occurrences in the corpus and topic t_j the ratio between word frequency in the topic and general word frequency in the corpus (known as Term Frequency/Collection Frequency Measure) normalized by the topic size (Eq. 2):

$$Tf \circ Cf(w_i, t_j) = \frac{freq_{tj}(w_i)}{freq_{corpus}(w_i)} * \frac{1}{size(t_j)} = \frac{count(w_i, t_j)}{\sum_l count(w_i, t_l) * \sum_m count(w_m, t_j)} \quad (2)$$

After scores are calculated for all words, we sort the words for each topic from their highest to lowest score. Then we assign a topic dependent threshold for each topic determined by a Receiver Operating Characteristic (ROC) analysis as described in 4.4. All words above the threshold belong to the distinctive word set (DWS) per topic. In additionally conducted experiments with the corpus this method has proven to outperform other alternatives, such as TF*IDF or Term Discrimination Models (Salton et. al., 1975).

transportation	finance & property	emotion
lane	loan	loss
boarding	payments	xox
tires	printing	beyond
flight	sander	childish
wheel	cheque	love
license	paypal	bitching
roads	discount	mentally
battery	invoice	soo
plane	price	stressed
exit	dollars	nerves

Table 2: Examples of topic distinctive words.

Table 2 illustrates examples of high-scored retrieved distinctive words from several topics. It becomes evident that words with high scores are related to a topic in our intuition or mental concepts. However, frequently used general words, such as pronouns, prepositions, and common nouns, do not receive high scores, because of their vast number of occurrences in other topics, e.g. "never", "flat", "boy", "you", or "from". Topics that are more descriptive or transactional in their orientation, such as "transportation" or "finance", generate better content distinctive word sets than the ones with relational intent, such as "emotion".

4.4 Topic Recognition by Word Sets

In order to determine optimal thresholds (see 4.3) and to analyze the coverage and distinctiveness of the word sets, we divide the corpus into a training batch (90% of all messages) and a test batch (10%). The training batch is used for the calculation of word scores as described in 4.3. By iteratively increasing the score threshold which defines a word set, we calculate per iteration the amount of dialogs from the test batch containing at least one word of the set, for dialogs annotated with the affiliated topic as well as for dialogs tagged differently. Consequently, ROC curves are created for all topics. This process is performed in a cross validation manner (10-fold).

Figure 8 shows the ROC curves for the topics "human relations", "activities & events", "finance & property", and "food & drinks", averaged over the 10-fold iterations.

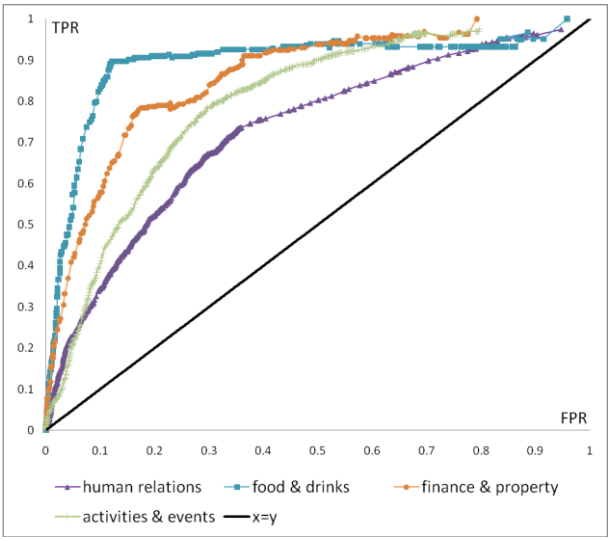

Figure 8: ROC curves for selected topics including best and worst performing topics with x axes for false positive rate (FPR) and y axes for true positive rate (TPR).

These results show that once appropriate thresholds are chosen, relatively small DWS, mostly ranging between 60-120 words per set, have the capability of achieving a true positive rate (TPR,

also known as recall) of 80.3% for topic dialogs with an average false positive rate (FPR, also known as fall-out) of 26.8%, even with a relatively naïve classification method. Table 3 provides detailed results of TPR and FPR. Topic DWS for more descriptive or transactional topics (e.g. "transportation", "food & drinks") manage to distinguish better than relational targeted topics, such as "emotion" and "human relations", since words like "love", "babe", or "thank" are highly related to the "emotion" topic, but also appear in many other topics. Hence, these words are increasing the FPR.

Eventually, the word sets chosen by optimal thresholds allow us to quantify topic recognition of dialogs. We automatically assign topics to each dialog in the corpus according to the described algorithm. Then we compare these topics to the manually annotated topics and measure recall and precision per dialog, denoted (Eq. 3):

$$recall = \frac{\#correct\_matched\_topics}{\#annotated\_topics}$$
$$prec = \frac{\#correct\_matched\_topics}{\#matched\_topics} \quad (3)$$

The average recall and precision rates over all dialogs are 73.5% and 44.3%, respectively. Taking into account the complexity of the recognition task due to the possibility of multiple topic assignment for each dialog, the results strengthen the hypothesis of the positively measureable interrelation between topics and wording.

4.5 Comparison to Full Vocabulary Models

Finally, we wish to better understand the impact of DWS, in comparison to the general language derived from the topic text, which is motivated by the fact that speech applications rely on SLMs. To this end, we construct a unigram language model binary classifier for each topic as baseline and perform a 10-fold cross validation classification task, to identify whether a given dialog is related to the topic or not, using the following formula (Eq. 4), where D_i is the i^{th} dialog and M_t is the language model of topic t:

$$topic^*(D_i) = \arg\max_{t \in topic, \overline{topic}} (D_i \mid M_t)$$
$$= \arg\max_{t \in topic, \overline{topic}} (\prod_{w \in D_i} p(w \mid M_t)) \quad (4)$$

Table 3 summarizes the results of TPR and FPR of the two approaches. As expected, the DWS approach suffers from a higher FPR, due to a lack of weights and relative comparisons to other classes. Since the differences in FPR between the two methods are not immense, we conclude that our chosen word sets are indeed distinctive, and with proper tuning have the potential of achieving better results. On the other hand, the DWS approach manages to outperform language models in terms of TPR. Hence, most of the information needed for the identification of dialog topics is provided by distinctive words to a significant higher extent as by the rest of the vocabulary.

Topic	DWS		Language models	
	TPR	FPR	TPR	FPR
Activities & events	81.9	34.7	64.1	22.8
Appoint. & schedule	69.5	31.0	82.6	21.4
Transportation	78.7	17.3	68.8	9.8
Finance & property	77.9	17.0	76.5	9.6
Food & drinks	88.4	11.7	74.1	10.6
School & work	80.9	22.4	54.3	14.0
Technology	92.4	28.7	75.5	12.6
Emotion	80.7	34.4	71.3	12.7
Human relation	72.2	34.7	69.8	20.8
	80.3	**26.8**	**70.7**	**14.9**

Table 3: True and false positive rates for all topics using DWS classification and language models.

5 Conclusion

The primary motivation of this study has been to estimate and facilitate the potential integration of contextual knowledge, in particular topics, into SMS dictation by voice. We have identified the interrelation between conversed topics and the choice of words in SMS dialogs as a key property, which needs to be quantified. After creating an annotated corpus and developing a classification method based on topic distinctive word lists, we have presented initial, promising results, which encourage further research.

Our study exposes also some challenges, which may not be easy to address. It would be useful to have a larger annotated corpus. Fully automated annotation of topics seems hardly achievable in view of our results. We may therefore rely on semi-supervised or unsupervised learning algorithms. Moreover, the study explores the relation of topics to single words. It needs to be enhanced

to phrases, because SMS dictation by voice relies on higher order n-gram SLMs.

In summary, when taking the next step and moving towards speech applications, we expect performance improvement after making topic knowledge useful for SMS dictation.

References

Agar, Jon (2003). Constant touch: A global history of the mobile phone. Cambridge, UK: Icon Books.

Aw, AiTi, Zhang, Min, Xiao, Juan & Su, Jian (2006). A phrase-based statistical model for SMS text normalization. In Proceedings of COLING/ACL, Sidney, AU.

Chafe, Wallace (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (Ed.), Subject and Topic (pp. 25-55). New York: Academic Press.

Cook, Paul & Stevenson, Suzanne (2009). An unsupervised model for text message normalization. In Proceedings of the NAACL HLT, Boulder, CO.

Fairon, Cédrick & Paumier, Sébastien (2006). A translated corpus of 30,000 French SMS. In Proceedings of LREC, Genova

Goggin, Gerard (2006). Cell phone culture: Mobile technology in everyday life. New York: Routledge.

Hecht, Ron M., et. al. (2009). Information Bottleneck based age verification. In Proceedings of Interspeech, Brighton, UK.

Kobus, Catherine, Yvon, Francois & Damnati, Geraldine (2008). Normalizing SMS: are two metaphors better than one? In Proceedings of COLING, Manchester, UK.

Kothari, Govind, Negi, Sumit & Faruquie, Tanveer A. (2009). SMS based interface for FAQ retrieval. In Proceedings of ACL, Singapore.

Lambert, Lynn & Carberry, Sandra (1992). Using linguistic, world, and contextual knowledge in a plan recognition model of dialogue. In Proceedings of the 14th International Conference on Computational Linguistics.

Lenhart, Amanda, et. al. (2010). Teens and mobile phones. From Pew Research Center http://pewinternet.org/Reports/2010/Teens-and-Mobile-Phones.aspx

Levinson, Stephen C. (1983). Pragmatics. Cambridge: Cambridge University Press.

Li, Charles N. & Thompson, Sandra A. (1976). Subject and topic. A new typology of languages. In Li, Charles N. (Ed.), Subject and Topic (pp. 457-490). New York: Academic Press.

McEnery, Tony, Xiao, Richard & Tono, Yukio (2006). Corpus-based language studies. An advanced resource book. London, New York: Routledge.

Meyer, Charles F. (2002). English corpus linguistics. An introduction. Cambridge: Cambridge University Press.

Molnár, Valería (1993). Zur Pragmatik und Grammatik des Topik-Begriffes. In Reis, Marga (Ed.), Wortstellung und Informationsstruktur (pp. 155-202). Tübingen: Niemeyer.

Pennell, Deana L. & Liu, Yang (2010). Normalization of text messages for text-to-speech. In Proceedings of ICASSP, Dallas, TX.

Salton, Gerard, Wong, Anita & Yang, Chung-Shu (1975). A Vector Space Model for automatic indexing. In Proceedings of Communications of the ACM, 18(11), 613–620.

Slonim, Noam & Tishby, Naftali (2000). Agglomerative Information Bottleneck. In Proceedings of NIPS 12.

Stutterheim, Christiane von (1997). Einige Prinzipien des Textaufbaus. Empirische Untersuchungen zur Produktion mündlicher Texte. Tübingen: Niemeyer.

Thurlow, Crispin (2003). Generation txt? The sociolinguistics of young people's text-messaging. From Discourse Analysis Online http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-01.html

Thurlow, Crispin & Poff, Michele (2009). The language of text messaging. In Herring, Susan C., Stein, Dieter & Virtanen, Tuija (Eds.), Handbook of the Pragmatics of CMC. Berlin and New York: Mouton de Gruyter.

Tishby, Naftali, Pereira, Fernando C. & Bialek, William (1999). The Information Bottleneck method. In Proceedings of 37th annual Allerton conference on communication, control and computing, Monticello, IL.

Wallis, Sean & Nelson, Gerald (2001). Knowledge discovery in grammatically analysed corpora. Data Mining and Knowledge Discovery, 5(4), 305-335.

Wu, Wei, Ju, Yun-Cheng, Li, Xiao & Wang, Ye-Yi (2010). Paraphrase detection on SMS messages in automobiles. In Proceedings of ICASSP, Dallas, TX.